No reference and reduced reference video quality metrics for end to end QoS monitoring

Patrick Le Callet, Christian Viard-Gaudin, Stéphane Péchard, Émilie Poisson Caillault

PAPER *Special Issue on Multimedia QoS Evaluation and Management Technologies*

No reference and reduced reference video quality metrics for end to end QoS monitoring

Patrick LE CALLET[†], **Christian VIARD-GAUDIN**[†], **Stéphane PÉCHARD**[†], *and* **Émilie CAILLAULT**[†], ***Nonmembers***

SUMMARY This paper describes an objective measurement method designed to assess the perceived quality for digital videos. The proposed approach can be used either in the context of a reduced reference quality assessment or in the more challenging situation where no reference is available. In that way, it can be deployed in a QoS monitoring strategy in order to control the end-user perceived quality. The originality of the approach relies on the very limited computation resources which are involved, such a system could be integrated quite easily in a real time application. It uses a convolutional neural network (CNN) that allows a continuous time scoring of the video. Experiments conducted on different MPEG-2 videos, with bit rates ranging from 2 to 6 Mbits/s, show the effectiveness of the proposed approach. More specifically, a linear correlation criterion, between objective and subjective scoring, ranging from 0.90 up to 0.95 has been obtained on a set of typical TV videos in the case of a reduced reference assessment. Without any reference to the original video, the correlation criteria remains quite satisfying since it still lies between 0.85 and 0.90, which is quite high with respect to the difficulty of the task, and equivalent and more in some cases than the traditional *PSNR*, which is a full reference measurement.

key words: *convolutional neural network, video quality assessment, MPEG 2, temporal pooling.*

1. Introduction

Objective video quality assessment means to compute automatically quality scores well correlated with the ones given by human observers. Such metrics can provide quality control of the compressed images and more generally Quality of Service (QoS) for image transmission and especially broadcasting. Image quality metrics can be divided in three categories :

- full reference metrics *(FR)*, which require the original image and the distorted image,
- reduced reference metrics *(RR)*, which require a description into some parameters of the original image and of the distorted image,
- and, no reference *(NR)* metrics, which only require the distorted image.

FR metrics have been intensively studied in literature. Ideally, they should be generic (suitable for any kind of distortions) and so allow coding schemes comparison. A particularly important issue in multimedia streaming application refers to in-service metrics with no-intrusive set-up, which allow to monitor and control systems while they are in operation. In such broadcasting purpose, with an emitter and a receiver, only *RR* and *NR* metrics are convenient for QoS monitoring since transmitting the whole reference image is not realistic at all. For such applications, *NR* metrics are the best choices since no extra data is added to the bitstream. The fact that neither the full reference video, nor some of its features are available for comparison makes an accurate assessment much more difficult. *RR* metrics represent good alternatives since the reduced reference can be coded and embedded in the bitstream and therefore constitutes a practical approach to quality evaluation, as long as the reduced reference size is not too large.

Other important issues in QoS monitoring are the complexity and the real time exploitation of the quality score. One may preferred to get one score for a long duration sequence (typically height seconds as in *FR* VQEG Test plan) while one may required to get quality score more often (one or two scores per second).

This paper presents metrics for color video quality assessment that provide quality values at a rate of two scores per second according to the data obtained from subjective tests under a SSCQE protocol with hidden reference removal and used for the performance assessment of the proposed metric. The video quality assessment system proposed in this work can be considered either as a *NR* metric or as a *RR* metric. In both cases, it corresponds to a very light system, allowing real time processing. In the *RR* case, since it is based on comparison of representations at both encoder and decoder side (only at the decoder side for *NR*), we have taken care to compute low complexity feature on the decoder side. Low complexity is not necessary at the encoder side since features can be computed off-line. Obviously, the *RR* system can be a much more accurate metric.

Section 2 gives a short overview of methods for objective quality assessment of videos that have been proposed in the literature. The proposed system is presented in section 3, whereas section 4 focuses on the neural network architecture. Section 5 reports on experimental results, demonstrating the method opera-

Manuscript received April 20, 2005.
Manuscript revised August 12, 2005.
Final manuscript received October 19, 2005.

[†]The authors are with IRCCyN, University of Nantes, France

tion under different conditions and for different input sources.

2. Related works

Peak Signal-to-Noise Ratio *(PSNR)* is currently widely used as a universal objective quality metrics. However, it is a pixel based fidelity metric, which does not always match well with the perceived picture quality. In the past decades, many objective quality metrics for measuring video impairments have been investigated [1]. Most of them used perceptual models to simulate the human visual system (HVS) and weight the impairments according to their visibility. Unfortunately, the HVS is so complex that existing perceptual models could not match to the real HVS well, and thus could not provide accurate rating of video quality. Another approach tried to exploit the properties of known artifacts, such as blocking artifacts, using feature extraction and model parameterization. This class of measure method focuses on the particular type of artifacts [2], [3] so it is normally more accurate than perceptual model based metrics. Due to QoS monitoring needs, such approach have been also proposed for *RR* and *NR* video quality metric like in [4] and [5]. However, it does not possess universality. Some authors intend to overcome this limitation using hybrid systems that involve feature extraction of coding artifacts and neural network systems capable of learning human perception as available from subjective tests. For example, a neural-based approach tested on VQEG (Video Quality Experts Group) test sequences is presented in [6]. This system, which is of *RR* type, process a 20-input feature vector that is forwarded to a radial basis function neural network (RBFNN) for classification. One limitation of this system is that the sequential nature of video is not taken into account, the RBFNN process data coming from a single frame, which is not very likely to correspond to HVS. With the work presented here, we extend the contribution of the NN, in order to mix the contribution of the features corresponding to successive frames. We have recently [7] presented the advantage of such an approach in the *RR* metric framework. In this paper, we not only extend the technique to the field of *NR* metric but also consider modification of the previous *RR* metric to address low complexity requirements for QoS monitoring.

3. The proposed video quality metric

Objective features are continuously extracted from video streams on a frame-by-frame basis; they feed the convolutional neural network estimating the corresponding perceived quality. In order to take into account color in our metric according to the human visual system, we use at the front end of the features extraction a decomposition into the three components

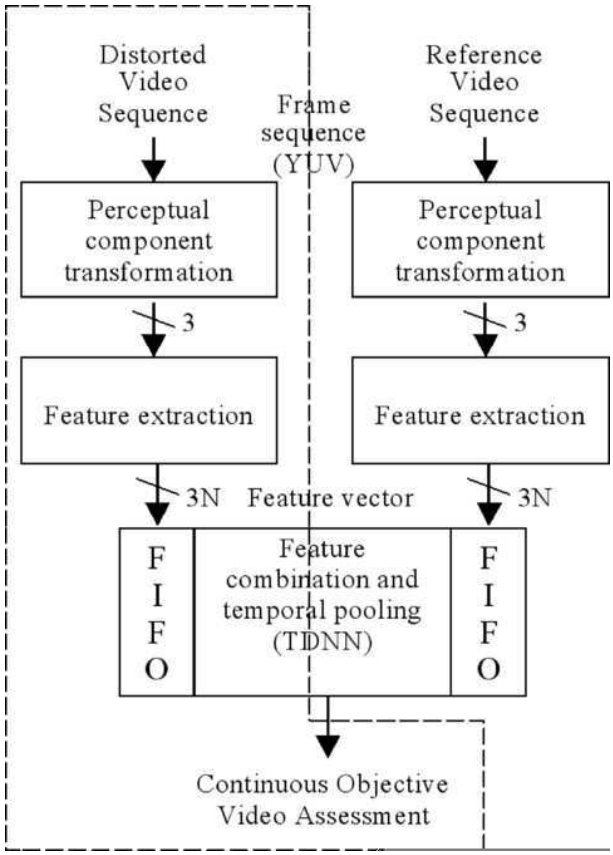


Fig. 1 Proposed video quality assessment system

of the Krauskopf's color space [8]. So, each feature is computed on each of the three components of the color space. Each feature represents a scalar value for the entire frame, so it corresponds to a full integration along the spatial dimension. As can be seen in Fig. 1, the same computations are done independently on both the distorted video sequence and the reference sequence, this is done in the case of the *RR* metric. But in the case of a *NR* metric, such a scheme allows to restrict quite easily the system to the left part of Fig. 1.

We have selected from the literature [9]–[11] a set of $N = 4$ features that are well suited in order to sum up the content of a frame. Voluntarily, these features have been chosen from models proposed for VQEG FR-TV I. They are probably not the best but our goal is to show that the way to pool such features can considerably improve the performance of a quality metric comparing to usual method (linear combination or Minkowski summation). Three of these features are totally content dependent (regarding frequency and temporal content). The last feature is more focused on distortion a priori related to blocking effect. Each of these 4 features is computed independently on the three perceptual components, they are presented in more details in the next sub-sections. Consequently, the global size of the fea-

ture vector describing every frame is at most $3 \times 4 = 12$ features. For the RR metric, in order to preserve low complexity at the decoder side, we have selected only one feature, so the global size of the feature vector is only three scalar value.

3.1 Frequency content features (GHV and $GHVP$)

The two first features, termed as GHV and $GHVP$, are derived from the work of [9], they represent the spectrum content of the videos. They have been previously elaborated to detect the blurring artifacts but are also sensitive to tiling distortions. These two features are computed from the two-dimensional histogram $SIH(r,q)$ where r is the magnitude of the gradient vector, and q is the orientation of the gradient vector with respect to the horizontal axis and $SIH(r,q)$ is the number of pixels in the gradient image whose gradient radius and angle is r and q, respectively.

The feature GHV whose value increases as the number or sharpness of horizontal and vertical edges increase is given as:

$$GHV = \frac{1}{p} \sum_{r} \sum_{\theta} SIH(r,\theta) \cdot r \tag{1}$$

with $0 < C_a \leq r \leq C_b$ and $\theta = \frac{k\pi}{2}, (k = 0,1,2,3)$ where r and q are as defined above, C_a and C_b are clipping limits and p is the number of pixels in the image.

In order to separate blurring from tiling, the $GHVP$ feature that characterizes the edge content of the image *without* the inclusion of horizontal and vertical edges is also computed:

$$GHVP = \frac{1}{p} \sum_{r} \sum_{\theta} SIH(r,\theta) \cdot r \tag{2}$$

with $0 < C_a \leq r \leq C_b$ and $\theta \neq \frac{k\pi}{2}, (k = 0,1,2,3)$.

3.2 Temporal content feature: Power of frame difference (P)

The next extracted feature, P, is derived from the work of [10]. They consider the following distortions: flicker, jadder, moving blurred images, random noise and edge jitter, and define linear combinations of some distortion factors using properties of visual perception. These combinations, which are explicitly defined in their work, are based on the power of the frame difference images computed respectively on the original and on the distorted video sequences. In our work, we will just keep the computation of the power of the frame difference and use it as an input feature for the NN. It will be the responsibility of the NN to model the distortions. The following computations are proceeded:

Frame difference:

$$d(t,m,n) = I(t,m,n) - I(t-1,m,n) \tag{3}$$

Power of frame difference:

$$P(t) = \sum_{m,n}^{all} \{d(t,m,n)\}^2 \tag{4}$$

3.3 Blocking measure (B)

This last measurement is mainly dedicated to exhibit blocking effects. It is based on the method proposed in [11], which has been recently simplified [12]. They apply 1-D FFTs to horizontal and vertical difference signals or rows and columns in the image to estimate the average horizontal and vertical power spectra. Peaks in these spectra due to 8×8 block structures are identified by their locations in the spectra. The power spectra of the underlying non-blocky images are approximated by median-filtering these curves. The overall blockiness measure, feature B, is then computed as the difference between these power spectra at the locations of the peaks. Integration of masking effects is possible with this scheme while it has not been used in our implementation.

4. Neural network architecture

The last stage of the system, presented in Fig. 1, corresponds to the feature combination and the temporal pooling of the feature vector sequence. Designing such a model is not straightforward. Ideally, approaches based on models of the human visual system (HVS) are the most general and potentially most accurate ones [13]. However, the HVS is extremely complex, and many of its properties are still not well understood today, specifically when one desires to take into account the temporal dimension of videos. Besides, implementing these models is computationally expensive due to their complexity.

To overcome these difficulties, we propose to base this function on a learning algorithm that will be capable of generalizing the observed behavior from a collection of subjective tests. As a trade-off between an explicit model that would require an in-depth knowledge of the HVS and a complete black-box system ignoring all the a priori knowledge, we introduced a neural net (NN) approach using a constrained architecture that is well suited to mimic the temporal integration of distortions. This architecture corresponds to a time delay neural network (TDNN), which performs convolution functions on the video sequence. It allows to model the following behaviors [14]: 1) assessor's reaction times are subject to delays; 2) time-consecutive frames tend to interfere with one another, and 3) the most recent frames of a sequence have a greater effect on the overall quality rating. The idea is to perform the same kind of computation at every place in the video stream based on a local receptive field. This is typically the principles involved with convolutional NN (CNN) [15]. In

our case, the convolution kernels will be defined along the temporal axis, leading to the so-called Time Delay Neural Network (TDNN). TDNNs are well suited to sequential signal processing [16]. They allow to preserve the sequential nature of data, in contrast with standard multilayer perceptron (MLP) where the topology of the input is entirely ignored. On the contrary, video sequences have a strong local structure: frames that are temporally nearby are highly correlated. Local correlations are the reasons for the well-known advantages of extracting and combining local features before processing temporal objects. With CNN, a given neuron detects a particular local feature of the video stream. It performs a weighted sum of its inputs followed by a non-linear squashing function (sigmoid). Its receptive field is restricted to a limited time window. The same neuron is reused along the time axis to detect the presence or absence of the same feature at different position of the video stream. A complete convolutional layer is composed of several feature maps, so that multiple features can be extracted at each temporal position. This weight sharing technique greatly reduces the number of free parameters and hence trained networks run much faster and require much less memory than fully connected NN. With local receptive fields, neurons can extract elementary visual distortions in videos. These distortions are then combined by the subsequent layers in order to detect high-order features.

In addition to the TDNN layers, the upper layers are standard fully connected layers. With this application, the last layer consists of a single neuron fully connected to the previous layer; the output of this neuron will be trained to estimate the Differential Mean Opinion Score *(DMOS)* value as it has been computed from human observers subjective scoring. A detailed view of the TDNN architecture is presented in Fig. 2.

The weight matrixes w, where indexes l, f, and t in Fig. 2 refer to layer, feature, and time position within the architecture, are learnt during the training step. It uses a standard stochastic gradient back-propagation algorithm adapted to respect the constraints of weight sharing [17]. The main change here is the computation of the local gradient of the backpropagated error signal with respect to the shared weights. Considering that every feature contains in fact a single neuron with multiple instances, the local gradient for this neuron is simply the summation of the local gradients over all instances of it [15].

From this general architecture, many parameters have to be defined to customize a specific learning machine. Different values for these parameters have been experimented, the selected values being given below:

- number of layers of the extraction sub-system (2),
- size of one layer with respect to the time axis ($T =$ 125 frames, i.e. 5s),
- size of one layer with respect to the feature axis ($nb\_feat = 20$),
- size of the convolution field with respect to the time axis ($field = 20$),
- temporal delay between two convolution fields ($delay = 5$),
- number of layers of the MLP sub-system (3),
- numbers of neurons of the hidden layer (50).

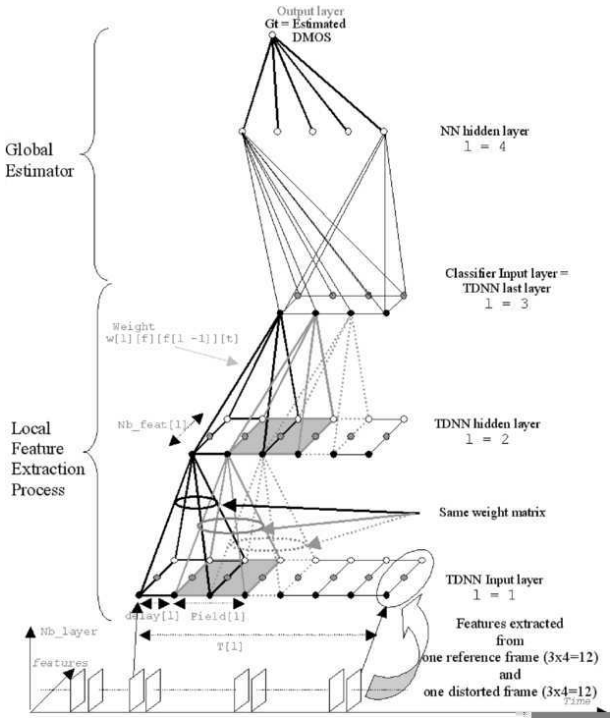


Fig. 2 Generic TDNN architecture

For example, the number of layers has been set globally to 4, including 2 layers for the local feature extraction sub-system, and 3 for the fully connected NN at the upper level, which correspond to one input layer — actually, the output layer of the TDNN subsystem, one hidden layer and an output layer with a single neuron. Time effect on the visual system is certainly difficult to model precisely. We assumed that beyond 5 seconds there is no significant contribution to the continuous perceived quality, it is the upper limit that our model enables. The NN has then the responsibility to weight accordingly the different frames embedded within this 5 second sequence. Consequently, the value of T, which refers to the number of frames involved in the computation of a score, has been set to $T = 5\text{s} \times 25\ \text{f/s} = 125$ frames.

5. Experimental results

5.1 Material available for training and testing

In order to train the TDNN we have used subjective quality assessment material provided by TDF. This material is composed of four reference videos of about

Video name	# video	Bit rates (Mbits/s)	# video scoring	Average $CI/2$ %
Cooking	4	2, 3, 3.5, 5	1440	13.7
Football	6	2, 3, 3.5, 4, 5, 6	2160	13.2
Horses	4	2, 2t, 3, 3t t = transcoded	1440	14.2
Road	3	2, 3, 6	1080	13.1

Table 1 Video sequences

three minutes long, (named Road, Cooking, Horses and Football), and 17 different distorted sequences, each of them corresponding to encoded and/or transcoded reference videos in MPEG2 at different rates (see Tab. 1 for details). For all of these video sequences, TDF has provided the corresponding subjective assessments results obtained with human observers. Subjective tests were running with 15 observers using a SSCQE protocol with hidden reference removal in normalized conditions and environment according to recommendations ITU-R BT.500-10. Subjective scores (MOS) consist of a quality rating sampled twice a second. It is easy to derive *DMOS* (difference of MOS between two conditions) with an associated Confidence Interval (CI) obtained according to subjective measurement procedures. In our case, the reference sequence has been produced by TDF using a MPEG codec at a bit rate of 8 Mbits/s which ensures a very high quality, although in some circumstances, impairments could be perceptible when compared with the uncompressed original videos. So, the system is fitted in order to compare quality loss between 8 Mbits/s and more severe rates.

To share this material between a training and a testing database, a separate video content is associated with the training and the testing video sequences. Considering the four sets of video contents (Road, Cooking, Horses and Football), sequentially, we trained the system on three out four of these sequences and tested it on the remaining sequences. For example: Football, Horses, Road (13 videos) compose the training set, and the test set is composed of the remaining group of videos, in this case: Cooking (4 videos). With this procedure, we are sure that the sets of images of the training set and test set come from disjoint video contents.

5.2 *PSNR* performance

Although *Peak-Signal-to-Noise-Ratio (PSNR)* is a *FR* indicator to assess the quality of reconstructed images, it has the advantage of being an easy and well known measurement to evaluate the performance of a compression technique. Table 2 provides the absolute value of the Linear Correlation Coefficient *(LCC)*, between *DMOS* and *PSNR* for the four testing databases. We produce two *PSNR* values per second by computing the mean of the obtained frame by frame *PSNR* for half a second.

Test database	Cooking	Football	Horses	Road
LCC $PSNR/DMOS$	0.915	0.863	0.875	0.788

Table 2 Full Reference $PSNR$ metric

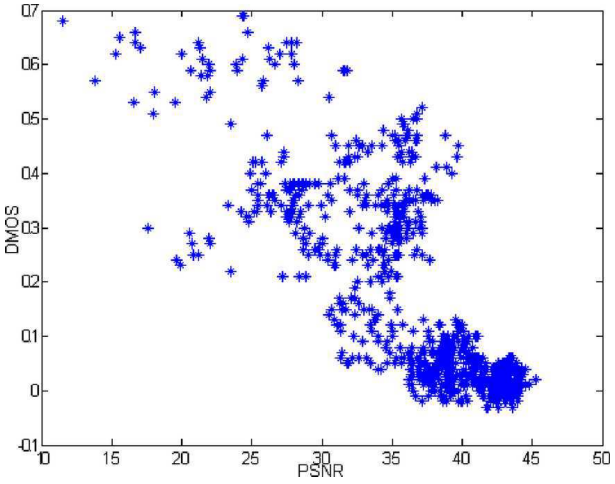


Fig. 3 $PSNR$ metric, scatter gram of Road videos

5.3 Quality assessment results

As a measure of performance of the proposed objective scoring method, three main indicators will be presented. One will be the root mean squared error on the test set, defined as:

$$J_{rmse} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} J_t} \tag{5}$$

where N is the number of scores computed on the test video sequences, and J_t is the network cost function, which is expressed as:

$$J_t = (DMOS_t - G_t)^2 \tag{6}$$

where $DMOS_t$ is the actual subjective score derived experimentally from the panel of observers and G_t is the output of the TDNN.

The second indicator is the Linear Correlation Coefficient *(LCC)*, which expresses the monotony between *DMOS* and objective scoring. The third one represents the percentage of the marks given by the TDNN that lays inside $\pm \frac{CI}{2}$ confidence interval margins (*Count*). For *PSNR*, it is only possible to compute *LCC*.

We have conducted two main sets of experiments. The first one is dedicated to evaluate the *NR* system, where only features extracted from the distorted videos are used, the number of which being 12 per frame ($3 \times 4 = 12$). Synthesized results are given in Tab. 3 for the four possible test sets, and a more detailed example is illustrated in Fig. 4 and 5. From Tab. 3, it can be

Test database	J_{rmse} %	LCC	$Count$ %
Cooking	11.3	0.81	79.6
Football	18.3	0.85	59.7
Horses	12.3	0.90	76.7
Road	9.2	0.85	89.8

Table 3 No Reference quality assessment, results on the test databases

observed that the mean quadratic error remains reasonable, except on the football sequence, specifically when we compare it with the CI values as given in Tab. 1. LCC values have to be compared with those obtained on the same test set with the traditional *Full reference PSNR* metric, and presented in Tab. 2. The proposed *No Reference* objective video quality assessment method clearly outperforms the basic *PSNR* metric for Road and Horses testing database. It is slightly equivalent for Football database while *PSNR* is clearly better for Cooking database.

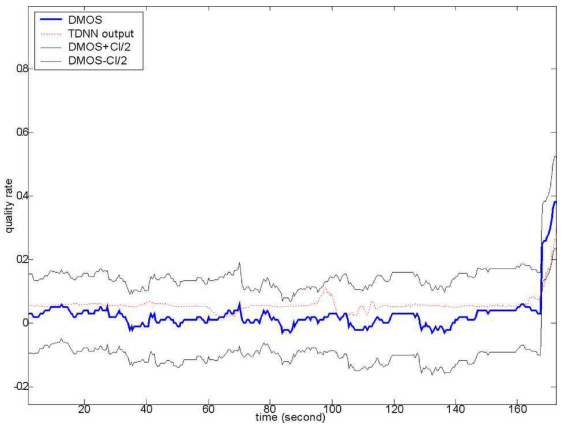
The results we obtained when the Road videos are used for testing, whereas Cooking, Foot, and Horses were used for training, are presented in more details in Fig. 4. It corresponds successively to three different bit rates: a) 6 Mbits/s, b) 2 Mbits/s, and c) 3 Mbits/s. The $DMOS$ are computed from the scores given by the human observers, and the $DMOS + \frac{CI}{2}$ and $DMOS - \frac{CI}{2}$ curves are also displayed, where CI represents the 95% Confidence Interval. The dashed-dotted line is the output of the TDNN, which represents the predicted score from the no reference features. In most of the situations, the predicted score lies inside the confidence interval range and follow quite well the human assessment. Except in the case of the most disturbed video at 2 Mbits/s, the predicted quality rate remains very close to the $DMOS$ subjective marks.

Chart presented in Fig.5 displays, for every scoring event (3 x 2 x 180 s = 1080) of the Road test set, on the x-axis the output of the TDNN, and on the y-axis the corresponding subjective scores *(DMOS)*.

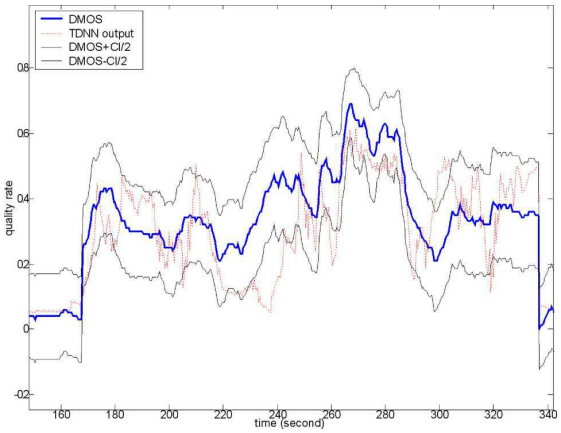
The second set of experiments has been carried out in the framework of the RR metric. In such a case, both features from the reference videos and the distorted videos are extracted.

In order to downsize the proposed system, and for an engineering standpoint, in order to guarantee its effectiveness in real-time production applications, we have studied and selected the best informative feature among the set of four, namely GHV, $GHVP$, P, and B. Table 4 shows the individual performance of each of the features on the overall video set. It can be concluded that the best performances was achieved using the P features alone. This is all the more appreciated result since P feature requires poor complexity.

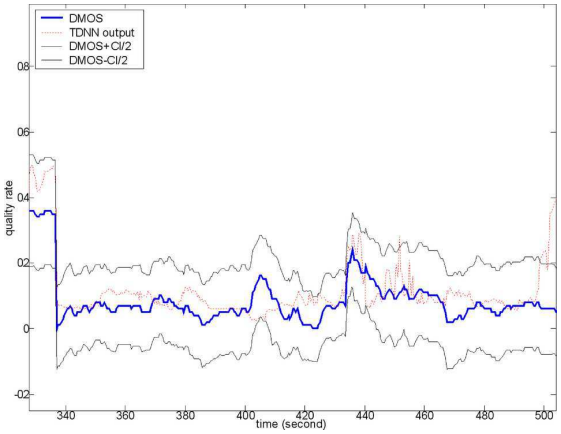
Corresponding results are displayed in Tab. 5. Once again, Football video gets poor results compared to the three other videos. Nevertheless, a significant improvement has been achieved and the linear correlation criterion ranges from 0.90 to 0.95. Moreover, the metric outperforms *PSNR* in all the case while it is RR based and equivalent in terms of complexity (slightly



(a) 6 Mbits/s



(b) 2 Mbits/s



(c) 3 Mbits/s

Fig. 4 *NR* metric, Road videos at three different bit rates

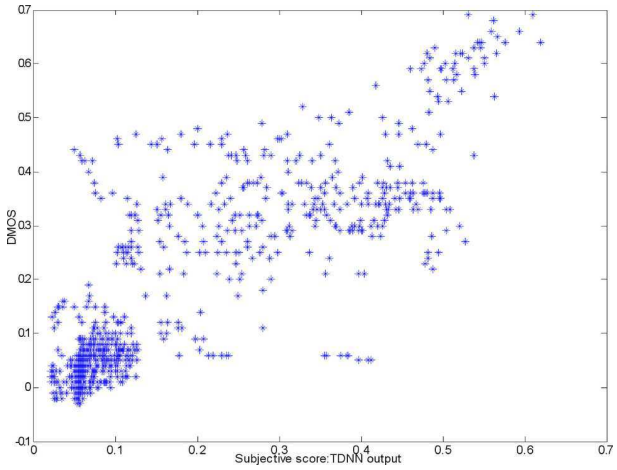


Fig. 5 *NR* metric, scatter gram of Road videos (1080 points)

Feature selected	J_{rmse} %	LCC	$Count$ %
GHV	11.9	0.83	73.1
$GHVP$	11.2	0.86	77.5
P	9.6	0.91	79.1
B	13.8	0.78	67.1

Table 4 Feature sensitivity w.r.t to Reference quality assessment, results on the global test database

Test database	J_{rmse} %	LCC	$Count$ %
Cooking	6.11	0.95	90.1
Football	12.9	0.94	61.6
Horses	6.8	0.95	87.6
Road	7.8	0.90	92.0

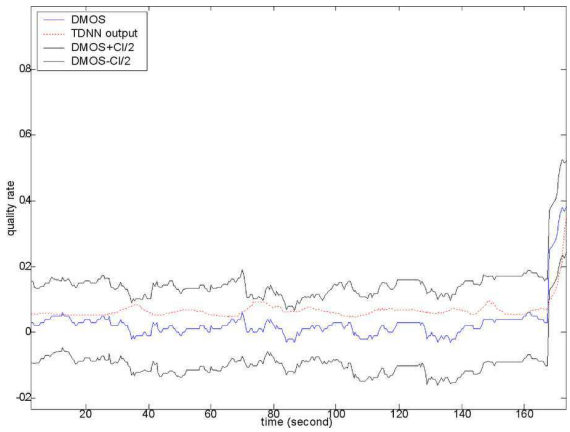
Table 5 Reduced Reference quality assessment, results on each video of the test databases

more due to the Neural Network propagation). The results we obtained when the four Road videos are used for testing (last row of Table 5), whereas Cooking, Football, and Horses were used for training, are presented in Fig. 6.

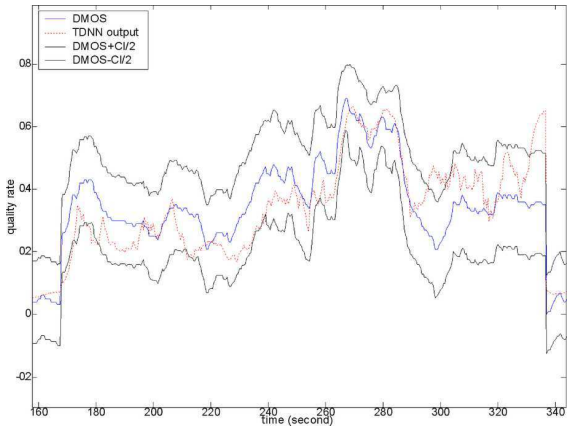
Figure 7 clearly illustrates the high correlation rate, which reaches 0.90, between the subjective *DMOS* values and the objective values computed with the proposed *RR* metric.

6. Conclusion

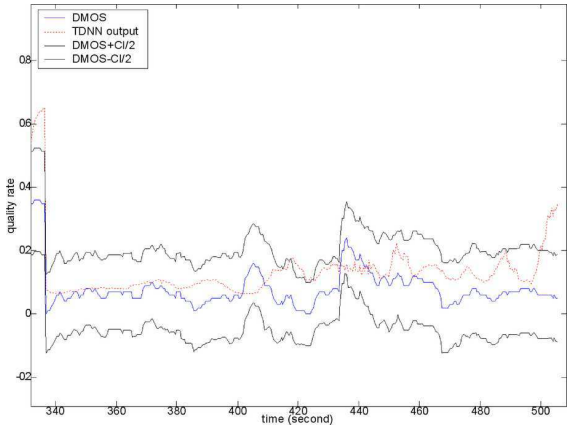
In this paper, we have demonstrated that TDNN can be used to asses the perceived quality of video sequences by realizing a non-linear mapping between non subjective features extracted on the video frames and subjective grades. More generally, it provides a method to combine features and to achieve temporal pooling. The method could be easily declined with other features. We have validated our approach using quite a large database that is composed of different video contents and different bit rates. Reference videos was based on a 8 Mbits/s bit rate whereas distorted videos have been produced with a bit rate varying from 2 to 6



(a) Road video, 6 Mbits/s



(b) Road video, 2 Mbits/s



(c) Road video, 3 Mbits/s

Fig. 6 *RR* metric, Road videos at three different bit rates

Mbits/s. On the test set, which was independent of the learning set, a linear correlation criteria ranging from 0.90 to 0.95 has been obtained between the output of

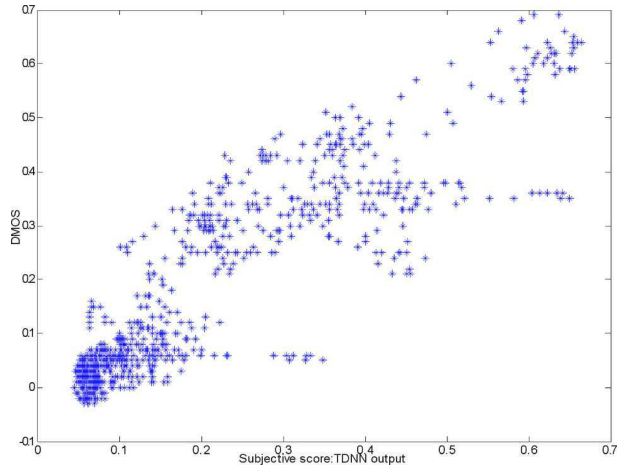


Fig. 7 *RR* metric, scatter gram of Road videos (1080 points)

the *RR* system and the subjective score provided by human observers. When we restrict the inputs to the features coming from the distorted videos in order to define a fully *NR* system, the correlation drops down to 0.85-0.90, it is still higher than the *PSNR* measurement, which was between 0.78 and 0.91. It shows a remarkable generalization capability of the neural network. The key factor of the proposed architecture relies on the set of convolutional neurons, which slides along the time axis sharing the same set of weights. They allow to perform the time integration function, which is not obvious to model and for that reason not always taken into account in other systems without introducing tremendous complexity. As a comparison, when a single frame ($T = 1$ instead of 125) is used to carry out the computation of the objective score, the results we got are really worst: the correlation was around 0.80 for the 4 test videos with the *RR* system using the P feature alone (to be compared with 0.90–0.95).

We have in mind to extend this system along two directions. One would be to take into account more general degradations than those due to lossy compression algorithms. Specifically, a complementary set of features sensitive to transmission errors has to be defined, and of course, for the training purpose, a new database including such kind of errors should be available. The second extension consists in replacing the spatial integration that is carried out during the feature extraction process by a learning stage that will be incorporated in the neural architecture. The same kind of approach, with convolutional neurons could be used. It leads to Space Displacement Neural Network (SDNN), which has already been used with success and combined with TDNN, for example in document image processing [17].

Acknowledgments

The authors wish to thank TDF for providing the databases used in the experiments related in this paper.

References

[1] A.B. Watson, J. Hu, and J.F. McGowan III, "DVQ: A digital video quality metric based on human vision," Journal of Electronic Imaging, vol.10, no.1, pp.20–29, 2001.

[2] H.R. Wu and M. Yuen, "A generalize block-edge impairment metric for video coding," IEEE Signal Processing Letters, pp.317–320, 1997.

[3] S.A. Karunasekera and N.G. Kingsbury, "A distortion measure for blocking artifacts in image based on human visual sensitivity," IEEE Transactions on Image Processing, vol.4, no.6, pp.713–724, 1995.

[4] J. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," Proc. SPIE Visual Communications and Image Processing, 2003.

[5] M. Farias, No-reference and reduced reference video quality metrics : new Contributions, Ph.D. thesis, University of California, 2004.

[6] S. Yao, W. Lin, Z. Lu, E. Ong, and X. Yang, "Video quality assessment using neural network based on multi-feature extraction," Proc. SPIE Visual Communications and Image Processing, ed. T. Ebrahimi and T. Sikora, pp.604–612, 2003.

[7] P. Le Callet, C. Viard-Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," First International Workshop on Video Processing and Quality Metrics for Consumer electronics, Phoenix, 2005.

[8] D.R. Williams, J. Krauskopf, and D.W. Heeley, "Cardinal directions of color space," Vision Research, vol.22, pp.1123–1131, 1982.

[9] D. Melcher and S. Wolf, Objective Measures for Detecting Digital Tiling. Document Number: T1A1.5/95-104, http://www.its.bldrdoc.gov, 1995.

[10] T. Yamashita, M. Kameda, and M. Miyahara, "An Objective Picture Quality Scale for Video Images (PQSvideo) – Definition of Distortion Factors," Visual Communications and Image Processing 2000, Proceedings of SPIE, pp.801–809, 2000.

[11] Z. Wang, A.C. Bovik, and B. Evans, "Blind Measurement of Blocking Artefact in Images," International Conference on Image Processing, pp.981–984, 2000.

[12] Z. Wang, H.R. Sheikh, and A.C. Bovik, "No-Reference Perceptual Quality Assessment Quality of JPEG Compressed Images," International Conference on Image Processing, 2002.

[13] S. Winkler, "Issues in Vision Modeling for Perceptual Video Quality Assessment," Signal Processing, vol.78, no.2, pp.231–252, 1999.

[14] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective Quality Assessment of MPEG-2 Video Streams by Using CBP Neural Networks," IEEE Transactions on Neural Networks, vol.13, no.4, pp.939–947, 2002.

[15] C.M. Bishop, Neural Networks for Pattern Recognition, pp.116–161, ISBN 0-19-853849-9, Oxford University Press, 1995.

[16] A. Waibel, T. Hanazawa, G. Hinton, K. Shikan, and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.3, pp.328–339, 1989.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, 1998.

Patrick LE CALLET holds a PhD in image processing from the University of Nantes (2001). Engineer in electronic and informatics, he was also a student of the École Normale Supérieure de Cachan. He received in 1996 his agregation degree in electronics. Associate professor at the university of Nantes, he is engaged in research dealing with the application of human vision modeling in image processing. His current centers of interest are image quality assessment, watermarking technique and saliency map exploitation in image coding techniques.

Christian VIARD-GAUDIN is a specialist of pattern recognition and learning machine. His main topics of interest are image processing are more specifically handwriting character recognition. He has supervised several research projects and PhD students in these fields and will be program chair of IWFHR'10. He is associate professor at the University of Nantes.

Stéphane PÉCHARD is a PhD student, the subject of his thesis is the conception of an objective quality criterion for high definition television systems. He experimented the time-delay neural network and determined parameterization for best results. He is also engineer in electronic and informatics from the École polytechnique de l'université de Nantes.

Émilie CAILLAULT is a PhD student, the subject of her thesis is the architecture and training of neuro-markovian system for online handwriting recognition. She was at the origin of the development of many convolutionnal neural networks, first applied to handwriting recognition and then to the image quality evaluation.